A. Experimental details

In this section, we provide information about the hyperparameters used for GS and IPR methods as well as further experimental results supporting the conclusions in the main part.

We always evaluated IPR using neighborhood size $k = 3$ as suggested by the authors. For this, we used balanced sets R and E obtained by sampling $\min(|R|, |E)$ points from each of them. The hyperparameters used for GS are adjusted to the specific experiment and discussed in the sections below.

A.1. Contrastive Learning

Mode truncation experiment The number of representations corresponding to each class c_t for $t = 0, \dots, 11$ in the training and holdout splits of both $\mathcal{D}_f$ and $\mathcal{D}_m$ are shown in Table 2 (middle rows). The set R was composed of representations corresponding to the first 7 classes $c_0, \dots, c_6$ from the training split, which amounts to 3514 points. The respective sizes of the sets E_t are shown in the right column of the table. The $\varepsilon(1)$ threshold evaluated to 0.05 and 0.18 in case of $\mathcal{D}_f$ and $\mathcal{D}_x$, respectively. Note that the value is constant across all values of t because it was always estimated on R using same random seed. In this experiment, we evaluated GS using $L_0 = 64$, $\gamma = 1/128$, $i_{\max} = 10$ and $n = 1000$.

Table 2. Number of representations corresponding to each class contained in the training and holdout splits of both $\mathcal{D}_f$ and $\mathcal{D}_m$ datasets (middle columns). The respective sizes of each set E_t used in our experiments is shown in the right column.

CLASS	TRAIN	HOLDOUT	E_t
0	670	666	666
1	690	625	1291
2	395	373	1664
3	706	684	2348
4	349	429	2777
5	409	377	3154
6	295	309	3463
7	296	312	3775
8	292	310	4085
9	311	293	4378
10	258	279	4657
11	331	345	5002

Next, we demonstrate how the variations in the component consistency and quality thresholds η_c, η_q, respectively, can be used to evaluate only components of certain minimum quality. Since the components in the mode truncation experiment have both high consistency and high quality, we deliberately corrupted the sets E_t to obtain more inconsistent and homogeneous components. Instead of adding all images of the class c_t to E_{t-1}, we sampled a subset of them of a randomly chosen size. The $\mathcal{P}, \mathcal{R}$ scores obtained by varying t and η_c, η_q are shown in Figure 12. In the left panel, we visualize the scores obtained at a constant $\eta_c = 0$ and $\eta_q \in \{0, 0.1, 0.3\}$. In the right panel, we instead fix $\eta_q = 0$ and vary $\eta_c \in \{0, 0.4, 0.6\}$. In both panels, we additionally plot reference $\mathcal{P}, \mathcal{R}$ scores (in gray) obtained for $\eta_c = 0.45$ and $\eta_q = 0.75$. We observe that the scores correctly decrease when considering larger threshold values.

Evaluating class separability In Figure 10, we show network quality $q(\mathcal{G})$ obtained when varying the distance threshold ε on both Siamese and SimCLR models on $\mathcal{D}_f$. We observe that the Siamese network, in addition to having 7 components containing more than 100 points for $0.1 \leq \varepsilon \leq 0.6$ (Figure 7), also achieves higher network quality. This means that these components also contain many heterogeneous edges, indicating that R and E are also well geometrically positioned.

A.2. Generative models

In this experiment, R always contained VGG16 representations of 50000 training data points from the FFHQ dataset, while E contained 50000 representations corresponding to the images generated by a trained StyleGAN model. The threshold $\varepsilon(10)$ was estimated to 28.10 for all ψ. As before, we evaluated IPR using neighbourhood size $k = 3$, while evaluated GS on 10000 randomly sampled points using $L_0 = 64, \gamma = 1/1280, i_{\max} = 100$ and $n = 1000$ for GS following the authors' recommendations except for the value of n. Initially, we tried running GS with $n = 10000$ and using all 50000 points but stopped the evaluation because of too slow computations (around 8 hours CPU time per truncation).

Varying sample size We used this large-scale experiment to perform time complexity and robustness analysis for varying number of samples contained in the sets R and E. For this, we additionally subsampled 10000, 5000 and 1000 samples from both R and E and calculated GeomCa global scores using $\delta = \varepsilon(10)$. In Figure 13, we plot the resulting $\mathcal{P}, \mathcal{R}$ scores (left panel) and $c(\mathcal{G}), q(\mathcal{G})$ scores (right panel) obtained on all sizes of R and E. For comparison, we additionally visualize GeomCA and IPR obtained using all 50000 points as in Section 5. We observe that GeomCA returns consistent results for all sizes except for the case of 1000 points where we obtained slight inconsistencies for truncations $0.4 \leq \psi \leq 0.6$.

In Tables 3 and 4, we report (CPU-based) time analysis of GeomCA obtained on truncation $\psi = 1.0$ corresponding to the varying sizes of the sets R and E as above. In Table 3 we report the sizes (cardinality) of the R and E sets given as inputs to GeomCA (left column) as well as the sizes of the obtained sparsified sets R' and E' (middle and right columns, respectively). In parenthesis, we report the time it took to sparsifty each of the sets. In Table 4 we report the

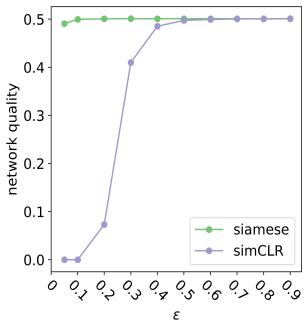

Figure 10. Network quality $q(\mathcal{G})$ (y-axis) obtained for Siamese and SimCLR models on $\mathcal{D}_f$ when varying distance threshold ε.

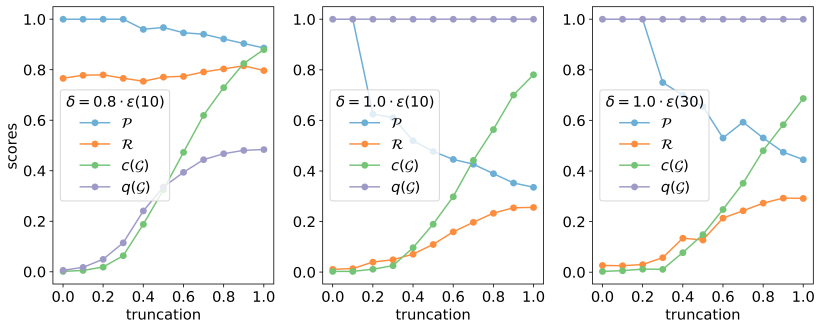

Figure 11. Precision $\mathcal{P}$, recall $\mathcal{R}$, network consistency $c(\mathcal{G})$ and network quality $q(\mathcal{G})$ obtained in the StyleGAN experiment when varying δ (left and middle) and when varying ε (middle and right).

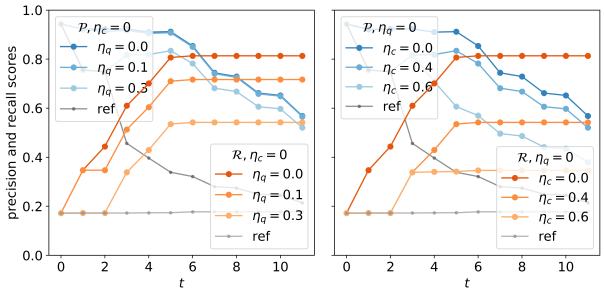

Figure 12. Precision and recall scores obtain when varying the consistency and quality thresholds η_c, η_q. Left: scores obtained for a fixed $\eta_c = 0$ and varying η_q. Right: scores obtained for a fixed $\eta_q = 0$ and varying η_c. Both panels show reference scores (gray) obtained using $\eta_c = 0.75$ and $\eta_q = 0.45$.

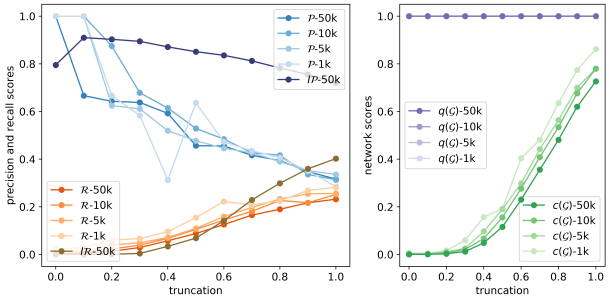

Figure 13. Results of StyleGAN truncation experiment obtained when varying the sizes of the sets R and E to contain 50000 ($50k$), 10000 ($10k$) 5000 ($5k$), and 1000 ($1k$) points. Left: GeomCA precision and recall scores $\mathcal{P}, \mathcal{R}$ as well as IPR scores obtained on 50000 points. Right: GeomCA network consistency $c(\mathcal{G})$ and quality $q(\mathcal{G})$ scores.

sizes of the vertex set $|\mathcal{G}|_\mathcal{V}$ and edge set $|\mathcal{G}|_\mathcal{E}$ of the resulting graph $\mathcal{G}$ build on the sparsified sets R' and E'.

Varying ε and δ parameters In Figure 11, we visualize $\mathcal{P}, \mathcal{R}, c(\mathcal{G})$ and $q(\mathcal{G})$ scores obtained when varying the sparsification parameter δ (left and middle), and distance threshold ε (middle and right). These results were obtained on R and E sets of size 5000. The middle panel corresponds to the parameters chosen in Section 5.We observe only slight changes in GeomCA scores when increasing ε from $\varepsilon(10)$ to $\varepsilon(30)$. On the other hand, decreasing δ results in more significant changes that, however, still reflect the correct structure of R and E. In particular, we observe that $\mathcal{P}, \mathcal{R}$ scores increase, while the network quality $q(\mathcal{G})$ decreases. This means that the connected components obtained when $\delta = 0.8 \cdot \varepsilon(10)$ contain more points than in case of $\delta = \varepsilon(10)$ but are also of lower quality. Note that the slight variations in network consistency are the result of applying sparsification with different parameters.

A.3. VGG16 Model

In Section 6,we defined two versions of the experiment where R and E sets contained 5 different classes of ImageNet datasets each. In version 1, we chose R to contain representations of images of classes *digital clock* (530), *espresso maker* (550), *frying pan* (557), *mixing bowl* (659) and *stove* (827), while E contained *Norwegian elkhound* (174), *Weimaraner* (178) , *Border terrier* (182), *golden retriever* (207), *Gordon setter* (214). In total, R and E contained 6258 and 6500 representations, respectively. In version 2, we randomly chose R to contain *Dungeness crab* (118), *shopping basket* (791), *lacewing* (318), *ski* (795) and *altar* (406), while E contained *fiddler crab* (120), *sliding door* (799), *sloth bear* (297), *beagle* (162) and *ladle* (618).

Table 3. Results of the sparsification applied to R and E sets of different sizes. In the middle and right columns, we show the size of the obtained sparsified sets as well as the elapsed time in parenthesis.

$\lvert R\rvert, \lvert E\rvert$	$\lvert R'\rvert$ [TIME]	$\lvert E'\rvert$ [TIME]
50000	9380 [1H 32MIN]	5349 [54MIN]
10000	2612 [3MIN]	1668 [2MIN]
5000	1387 [50S]	888 [31S]
1000	390 [2S]	295 [2S]

Table 4. Size of the vertex and edge sets obtained when building ε-graph on the sparsified sets R' and E' of sizes shown in Table 3. The elapsed time when building the graph is shown in parenthesis.

$\lvert R\rvert, \lvert E\rvert$	$\lvert R'\rvert$	$\lvert E'\rvert$	$\lvert\mathcal{G}\rvert_{\mathcal{V}} + \lvert\mathcal{G}\rvert_{\mathcal{E}}$ [TIME]
50000	9380	5349	18734 [11MIN]
10000	2612	1668	5468 [40S]
5000	1387	888	2930[11s]
1000	390	295	863 [1S]

In total, R and E each contained 6500 representations. The $\varepsilon(10)$ threshold was estimated to 196.34 and 197.59 in version 1 and 2, respectively. Moreover, we used neighborhood size $k = 3$ for IPR and $L_0 = 64, \gamma = 1/128, i_{\max} = 100$ and $n = 1000$ for GS.